Self-organizing Moral Systems:
Beyond Social Contract Theory[1]

"But what if morality is created in day-to-day social interaction,
not at some abstract mental level?"
~Frans de Waal

Gerald Gaus[2]

1 INTEGRITY AND RECONCILIATION IN MORAL THINKING

Contemporary moral and political philosophy is torn between two modes of moral reasoning. A common view of moral thinking — perhaps most characteristic of moral philosophy — understands reasoning about moral claims to be, in a fundamental sense, akin to reasoning about ordinary factual claims. On this commonsense approach, when Alf deliberates about a moral claim or demand (say, that people ought to respect property), he considers the best reasons as he understands them for and against the claim, including what he takes to be the correct normative principles, perhaps checks his conclusions with others to see if he has made any errors, and then comes to the conclusion, "we all ought to respect property." His moral reasoning may refer to facts about other people (say, their welfare), but it is not a general requirement on the moral reasoning of any competent agent that he always takes as one of his inputs the moral deliberations of others.

As far back as Hobbes (and probably a good deal further), many political philosophers have been deeply impressed how this first, individualistic, mode of moral reasoning leads to disagreement. When we employ our "private reason" there is, says Hobbes, great dispute about the application of the laws of nature, and so we require some common reason — a public reason — to reconcile our judgments and so provide a common interpretation of what the law requires.[3] More recently, public reason moral and political theory has focused on how reasonable disputes about the good and the right can be reconciled via common

[1] Versions of this paper were presented to the 2016 PPE journal conference, the 2017 PPE Society Conference, the Kadish Center Workshop in Law, Philosophy, and Political Theory at UC/Berkeley and at the University of Southern California. My thanks to all the participants for their helpful (and often spirited) comments and suggestions. [Forthcoming in *Philosophy, Politics and Economics*]

[2] James E. Rogers Professor of Philosophy, University of Arizona, Tucson AZ, 85721, USA. Email: jerrygaus@gmail.com.

[3] Hobbes, *Leviathan*, edited by Edwin Curley, (Indianapolis, IN: Hackett, 1994), p. 98 (chap. 15, ¶30). See nalso David Gauthier, "Public Reason" in *Public Reason*, Fred D'Agostino and Gerald F. Gaus, eds. (Brookeville, VT: Ashgate, 1988): 43-66 at pp 50ff. This same point was made earlier, and in more detail, by E.W. Ewin, *Virtues and Rights: The Moral Philosophy of Thomas Hobbes* (Boulder, CO: Westview, 1991), chap. 2. See also my "Public Reason Liberalism" in *The Cambridge Companion to Liberalism*, edited by Steven Wall (Cambridge: Cambridge University Press, 2015): 112-40.

conceptions of justice or shared moral rules.[4] This mode of moral reasoning aims to cope with — often by seeking to rise above — the moral disagreements engendered by the individual mode.

Yet many resist this second, Reconciliation, mode. "Why," they ask, must I to give up my own conclusions about justice for the sake of agreement? Isn't the search for reconciliation a violation of my moral integrity?" To this, the theorist of public reason reasonably responds by saying that *one's own understanding of morality* — that arrived at by the individual mode of reasoning — itself drives one to seek moral reconciliation with others, for on many matters one seeks to live a shared moral life with them. After all, one cannot institute a just system of property rights on one's own; one needs others, and this very moral need drives one to reconcile with others about the demands of justice. I shall argue that this response, as far as it goes, is correct: even when we consult only our own, individual, moral reasoning, almost all of us come to the conclusion that a moral life is to a very large extent a social achievement, and so we have moral reasons to seek out others who can share it with us. This, we shall see, is the great insight of the social contract. However, I shall argue, the contractual project flounders on the very supposition of moral disagreement on which it builds. It insists that because we disagree about the demands of morality and justice we need to reconcile with others, yet this very disagreement in individual moral reasoning leads us to disagree on a further matter: to what extent should we reconcile with others? Having reasonably seen that we disagree in our first-level moral judgments, the contractualist seeks to resolve this problem by assuming we agree on the required degree of reconciliation, and then constructs a device to secure it. The reflective contractualist may admit that this is a strong assumption but, she insists, one that *must* be made if we are to secure a social existence based on justice.

Having analyzed this problem in sections 2 and 3, sections 4 and 5 develop models of moral self-organization that show how it might be overcome. Each person, we shall assume, is committed solely to her own integrity — understood as her own judgments about first-level moral matters *and* the degree of reconciliation that her moral perspective endorses. I shall assume that free, reasonable and competent moral agents will disagree about both of these. Nevertheless, we shall see that under a surprising array of circumstances they can secure a shared moral life endorsed by all. The social contract's aim of full reconciliation can be secured without any contract device. And, perhaps most surprisingly, moral agents can

[4] This was the aim of my *Order of Public Reason* (Cambridge: Cambridge University Press, 2011). Although Rawls began by focusing on disputes arising from the good, by the introduction to the paperback edition of *Political Liberalism* he was explicit that individual reasoning about morality and justice also leads to disagreement. *Political Liberalism*, expanded edition (New York: Columbia University Press, 2005), pp. xxxv-lx.

arrive at this full reconciliation *just because* they disagree about the importance of reconciliation.

2 THE TWO MODES ANALYZED

2.1 The Individual Mode

To be a bit more precise, we can identify the individual mode of reasoning as:

> *The "I conclude we ought" implies "I ought" View:* As a competent moral agent, if (*i*) Alf conscientiously deliberates and concludes that, given what he takes to be the correct normative premises and relevant empirical information, one ought to ϕ (ought not ϕ, or may ϕ)[5] under conditions *C*, where this does not require taking account of the conclusions of the deliberations of others and (*ii*) he reasonably concludes that morality instructs that we all ought to ϕ under conditions C, then (*iii*) he ought to ϕ in circumstances *C*, even if others fail to do as they ought.

It is important that on the *"I conclude we ought" implies "I ought"* View (henceforth simply the "I conclude we ought" View), Alf does not typically assert that we all ought to ϕ in *C because he* has concluded that we ought to ϕ: Alf may believe that "we ought to ϕ" in *C* because it is a moral truth that we ought to ϕ, or that an impartial spectator would approve of our ϕ-ing. The important point is that once Alf conscientiously comes to the belief that one ought to ϕ in *C* — it is, we might say, his best judgment about the morally best thing to do — then, as a competent moral agent, he will justifiably ϕ in circumstances *C*, and indeed insist that we all do so, for that is what we ought to do.[6] And, as I have stressed, none of this necessitates (though it may be epistemically recommended) that Alf factors into his moral deliberation the moral conclusions of others.

2.2 The Reconciliation Mode

On one reading the social contract offers another view of justice and morality. Hobbes, Locke, Rousseau and Kant all hold that individuals' "private judgments" about morality or justice radically diverge, and because of this individual private judgment is an inappropriate ground — or at least, I shall argue, an inappropriate *sole* ground — for demands of justice. Kant famously insists that, even if we imagine individuals "to be ever so good natured and righteous," when each does what "seems just and good to him, *entirely*

[5] Henceforth I shall not state these alternatives, assuming that they are implicit.

[6] I assume here that conditions *C* are so defined that typical justifications for not ϕ-ing (duress, etc.) would show that *C* was not met. Recall that *C* cannot include the deliberations of others about what is moral in this circumstance. It can, though, take account of what Alf expects others to *do*. Given this, it is possible to construct a statement of conditions *C* that partially mimics the Reconciliation View, where Alf reconciles his understanding of morality with what others do, but he will still not reconcile with what others hold to be moral or just. I consider this possibility in §2.3.

independently of the opinion of others" they live without justice.[7] This apparently paradoxical conclusion — that a world of people who acted only on their own sincere convictions about justice would live without justice[8] — derives from two commitments of social contract theory. (*i*) It is taken as given that reasoned private judgments of justice inevitably conflict. This is partially because of self-bias, but only partially: innate differences in emotional natures, differences in beliefs that form the basis of current deliberations, differences in education, socialization and religious belief — all lead to pervasive disagreement. (*ii*) Secondly, it is assumed that a critical role of justice in our social lives is to adjudicate disputes about our claims and so coordinate normative and empirical expectations. For Kant the problem of universal private judgment was that "when there is a controversy concerning rights (*jus controversum*), no competent judge can be found."[9] Each, thrown back on her own reasoning, ends up in conflict, and ultimately unjust relations, with others. Understood thus, a necessary role of justice (or morality) is to provide an interpersonally endorsed adjudication of conflicting claims.[10] Securing justice, on this second view, is inherently something we do together.[11] If no other good-willed and conscientious moral agent accepts that in circumstances *C* justice demands ϕ, Alf's demand will not secure just social relations. Given points (*i*) and (*ii*), social contract theorists have endorsed a *Reconciliation Requirement*: individuals must seek out some device of agreement that reconciles their differences so that they can share convictions about what justice demands we do.

2.3 Justice as a Social Good

The worry about relying exclusively on "I believe we ought" reasoning is that my conclusion is about the justice of a joint action — what we do — but as a competent moral agent my deliberations control only what I do, not what others do, so I alone cannot produce the joint action. Consequently my "I believe we ought" judgment is very often ineffective in securing what I believe we both ought to do.[12] Rather than focus on individual

[7] Kant, *The Metaphysical Elements of Justice*, 2nd edition, edited and translated by John Ladd (Indianapolis: Hackett, 1999), p. 116 [§43]. Emphasis added.

[8] I have defended this paradox in some depth in "The Commonwealth of Bees: On the Impossibility of Justice-through-Ethos," *Social Philosophy & Policy*, vol. 33 (2016): 96-121.

[9] Kant, *The Metaphysical Elements of Justice*, p. 116 [§43]. Emphasis added.

[10] See John Rawls, "An Outline of a Decision Procedure for Ethics" in *John Rawls: Collected Papers*, edited by Samuel Freeman (Cambridge, MA: Harvard University Press, 1999): 1-19.

[11] It is not only contract theorists who think this. See G. A. Cohen, *Rescuing Justice and Equality* (Cambridge, MA: Harvard University Press, 2008), pp. 175ff.; R. B. Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), chap. 9.

[12] If, as do some, we suppose that judgments of justice are not intended to be action guiding, this is not a problem. I consider the extent to which judgments of justice are inherently practical in *The Tyranny of the Ideal: Justice in a Diverse Society* (Princeton: Princeton University Press, 2016), pp. 11-18. Here is it simply

actions, I shall henceforth consider disputes about the rules of justice that should be followed in some circumstance, supposing that rules are specific enough to provide people with guidance about how to act both now and in the future.[13] Alf, then, has concluded that, say, "we ought to both act on rule R_1," but the question remains whether he ought to do so if Betty refuses to. In this case, whether or not Alf's "I believe we ought" judgment is action guiding even for Alf depends on how he ranks the alternative joint rule-based outcomes in terms of justice. Contrast, for example, the interactions modeled in Figures 1 and 2.

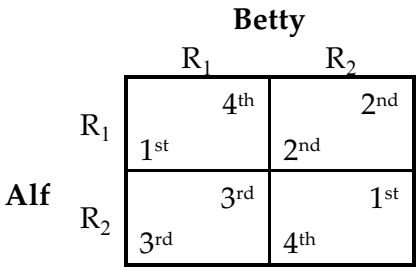

Figure 1: "I believe we ought ϕ" Implies "I ought to ϕ" (Go Your Own Way Game)

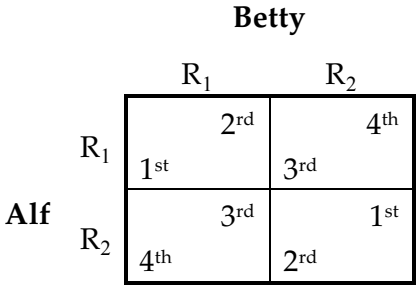

Figure 2: "I believe we ought ϕ" Does Not imply "I ought to ϕ" (Reconciliation Game)

In the interaction of Figure 1, each orders the outcomes: (1) we act on (my view of) the just rule; (2) I act on the correct rule of justice and the other acts on an inferior rule; (3) I act on the inferior rule and the other acts on the correct rule (at least someone does!); and (4) we both act on the inferior rule. In this game the sole equilibrium is that Alf acts on his view (R_1), and Betty acts on her view (R_2), of justice. At either of the coordination solutions (when both play R_1 or both play R_2), one of the parties would do better by changing his or her

assumed that our concern is action-guiding judgments; if there are other notions of justice, they raise different issues.

[13] See my *Order of Public Reason,* chap. 3 for a defense of these assumptions.

move, and acting on his or her favored interpretation of justice. So here even if the other does not do as you have concluded "we" ought, you still ought to do it. In this game each goes their own way.

Figure 2 does not support this conclusion. There are two equilibria in this impure coordination game: both act on R_1 and both act on R_2. Here both parties agree that from the perspective of securing just social relations, it is best that they do not necessarily act on their understanding of the best rule of justice, for if they do so they may secure social relations that are inferior from the perspective of justice. Thus, based solely on what they each believe justice requires, they are playing a Reconciliation Game. Now insofar as the aim of justice is securing social relations of a certain moral quality, we would expect Figure 2 would be a typical interaction. We need each other's cooperation to produce just social relations. It might be thought, though, that it would be enough for Alf and Betty to coordinate their *actions*, not the endorsement of the same rule of justice. So long as say, in Figure 2, Alf and Betty both act in the way R_1 requires, or R_2 requires, it does not matter what rule they endorse as just (enough). Justice, it might be thought, demands only coordinated action, not judgments. But while mere coordinated action would secure them some social goods, as Kant insisted they remain in conflict about justice. If, say, Betty acts on R_1 only because she is being coerced or simply to secure some non-moral social good, she will fail to see their joint action as just. Indeed while Alf thinks they are acting justly, Betty may well conclude that she is the victim of injustice, and will respond to the R_1 action as a wrong. Thus Alf is unable to appeal to the justice of R_1 to regulate their interactions, for she sees it as an unjust way of relating. Such an appeal would only aggravate conflict, not help resolve it. In such situations relations of moral accountability are undermined: Betty will not see herself as accountable for failing to abide by R_1, as she does not understand it to be a rule of justice.[14] They thus will have failed to secure an interpersonal relation informed by justice.

I suppose in this essay, then, that judgments of justice are typically about interactions along the lines of Figure 2 rather than 1. Judgments of justice seek to secure a certain type of moral relation and, especially among large groups of people, an individual's unilateral action seldom can secure this relation. It is this sense in which, I suppose, justice is a social-moral good. As Plato stressed in the *Republic*, it is about relations among individuals rather than unilateral action; that is why the theory of justice has been a part of social and political philosophy right from the beginning. Much moral thinking can be described as simply "I believe *I* ought" reasoning, where my only concern is what I ought to do, come what may. On can be chaste, honorable and honest alone; one cannot make promises, keep contracts, or determine mutual expectations about what is proper and improper on one's own. Here the moral theorist switches from "I believe I ought" to "I believe *we* ought" judgments, but once that move has been made, there is still a problem: *I* have concluded what *we* ought to

[14] See my *Tyranny of the Ideal*, pp. 180ff and *The Order of Public Reason*, chap. 4.

do, but I cannot secure this without your cooperation. That is why with justice we often play the Reconciliation Game in Figure 2 rather than the Go Your Own Way Game of Figure 1. This is the great insight of the social contract tradition.

3 RECONCILIATION VIA THE SOCIAL CONTRACT

We thus come to appreciate that determining the rules and institutions of justice is — at least to some extent — a social problem. A natural interpretation of this, which we might see as definitive of the social contract tradition, is to understand it as a collective problem. My concern here is not to present a thorough criticism of this idea — which, I think, has offered, and continues to offer, fundamental insights[15] — but to clearly contrast it to the alternative I shall develop.

3.1 Substantive Contracts: "We Believe We Ought" Views

Consider first a *substantive* version of the social contract, presenting a theory that suitably-characterized individuals would agree to common principles of justice to structure their social and political lives. Rawls's contractualist theory is, of course the quintessential case. Such substantive theories seek a "We believe we ought" judgment. The theory constructs an account of what reasonable and rational persons with certain motivations *would* agree what we should all do. The theorist, then, actually presents something like an "I believe that [we believe we all ought to]" view of justice. That is, the theorist provides an analysis of what *she* thinks *we* would collectively agree to as common, shared, principles of justice to regulate *our* relations. Rawls suggests an even more complicated view. As he presents his theory of justice, "you and I" develop a theory of what all reasonable people would agree to.[16] We thus seem to have something like an "I [Rawls] believe that [<you and I believe that> we believe we all ought to]" view.[17]

Now from the perspective of reconciliation, a theory that acknowledges that we disagree about justice, yet need to coordinate, is certainly a great improvement upon simple "I believe we ought to" reasoning, as it seeks to confront at a basic level the fundamental moral insight that unless you and I concur about the demands of justice, our social relations will be deeply flawed from the perspective of justice itself. Yet, at the end of the day, it is a theory of what the theorist believes that we all believe what we all ought to do. That is, at

[15] For important recent contributions, see Michael Moehler, *Minimal Morality* (New York: Oxford University Press, forthcoming); Peter Vanderschraaf, *Strategic Justice* (New York: Oxford University Press, forthcoming).

[16] John Rawls, *Political Liberalism*, pp. 28ff. See also Rawls, *A Theory of Justice,* revised edn (Cambridge, MA: Harvard University Press, 1999), p. 44.

[17] It is for this reason that Habermas is correct to characterize Rawls's theory as "monological" at the highest level. Jürgen Habermas, "Reconciliation Through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," *The Journal of Philosophy*, vol. 92 (March, 1995): 109-31 at p. 117.

the end of the day, it is one person's conviction about what we all believe we ought to do; and for the same reasons we disagree in our simple "I believe we ought" judgments, we disagree in our "I believe we believe we ought" judgments.[18] Rawls, indeed, came to recognize that reasonable people do disagree about the most reasonable conception of justice — about the conception of justice that we would agree to.[19] Rawls's own theory of "justice as fairness" — a theory about what we will agree on — is highly controversial. And even if two philosophers, Alf and Betty, accept Rawls's principles of justice, they are almost certain to disagree on their interpretation, leading them to interactions along the lines of Figure 2.

3.2 Bargaining about Reconciliation

The substantive social contract, then, seeks a path to reconciliation by articulating a collective judgment about justice — what *we* believe *we* ought to do — but it is ultimately a theorist's judgment about what is the collective judgment. Implicit in this is the theorist's judgment about how much reconciliation among diverse views is called for — but that is one of our deep disagreements, thus the substantive view struggles with the problem of disagreement of private judgments about justice.[20] If the problem is that each individual has a different estimate of the moral costs and benefits of reconciliation, a second version of the social contract seems promising. Combining substantive and procedural elements, the *Bargaining Contract* takes this problem seriously, seeking a reasonable balance between the two modes of reasoning. Abstracting from important technical details, the heart of this approach is to identify two points for each individual, say Alf and Betty: the justice-based "payoff" that one would receive from unilateral action based on one's "I believe we ought" reasoning (the so-called "no agreement point") and the best "payoff" one could receive from coordinated (reconciliation) action. Note, then, that the outcome depends on the relation between these two concerns for each agent. We suppose both would gain "moral utility" (better moral outcomes)[21] through some form of coordination; if this was not the case, they would not be playing the Reconciliation Game, but Go Your Own Way. Consider, then, Figure 3, which is an asymmetric version of the Reconciliation Game (Figure 2) in cardinal utility.

[18] Or "I believe that [<you and I believe that> we believe we all ought to]" judgments.

[19] Rawls, *Political Liberalism*, p. xlvii.

[20] See further my "Public Reason Liberalism."

[21] It is critical to keep in mind that "utility" is simply a mathematical representation of a person's ordering of states of affairs, not itself a good which is sought. To say that an agent maximizes utility is simply to see her as one whose actions are directed to obtaining the state of affairs that she ranks as best. If her rankings are based solely on moral criteria, then, assuming that common formal consistency conditions are met (e.g., if α is better than β, then β is not better than α), her choices can be represented in terms of maximizing (moral) utility: she has a moral utility function.

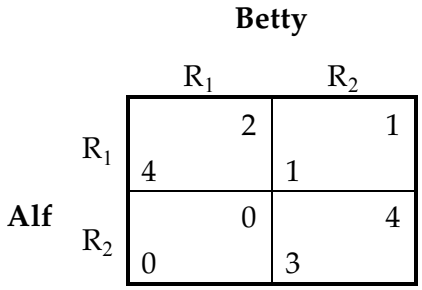

Figure 3: An Asymmetric Reconciliation Game

In Figure 3, the no-agreement point (R_1, R_2) is each person's third option (1, 1); it is the best outcome that each could receive if, as it were, they walked away from an agreement to coordinate. Both would gain by moving to (R_1, R_1) (4, 2) or to (R_2, R_2) (3, 4): the question is which is to be chosen. Drawing on bargaining theories that focus on division of a distributable good such as, say, money or time, some social contract theorists suggest we might see Alf and Betty's problem as deciding how much utility it would be rational for each to give up to secure a bargain. On the most common approach, axioms are defended that identify a unique division of the utility gains.[22] If we employ, for example, David Gauthier's principle of minimax relative concession, at (R_1, R_1) Alf makes no concession and Betty concedes 2/3; at (R_2, R_2) Betty makes no concession and Alf concedes 1/3.[23] Thus of the two, (R_2, R_2) minimizes the maximum relative concession. On Gauthier's account (R_2, R_2) is a more rational bargain: the relative concession of the party who gives the largest relative concession is less than in the (R_1, R_1) bargain.[24]

[22] See, for example, John Nash, "The Bargaining Problem." *Econometrica*, vol. 18 (1950): 155-62; David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986), chap. 5; Ehud Kalai and Meir Smorodinsky, "Other Solutions to Nash's Bargaining Problem," *Econometrica*, vol. 43 (1975): 513-18; Ehud Kalai, "Proportional Solutions to Bargaining Situations," *Econometrica*, vol. 45 (1977): 1623-30. For a noncooperative formulation, see Ariel Rubinstein, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, vol. 50 (1982): 97-109.

[23] According to minimax relative concession, we compute relative concession according to the formula:

$$\frac{u(fp) - u(fc)}{u(fp) - u(ip)}$$

where *u(fp)* is one's "first proposal" (the most which one could claim from the bargain without driving the other party away [for both this is 4]; *u(fc)* is what one actually receives from a bargain [at (R_1, R_1) this is (4, 2), at (R_2, R_2) this is (3, 4)] and so represents one's "final concession"; I suppose that *u(ip)* is the utility of one's initial position — the utility one comes into the agreement assured of, in this case (1, 1).

[24] There is a better solution I shall not explore: if they employ a correlated equilibrium and play (R_1, R_1) one-third of the time, and (R_2, R_2) two-thirds of the time, each concedes only 2/9. On correlated equilibrium, see my *On Philosophy, Politics, and Economics* (Belmont, CA: Wadsworth, 2008), pp. 140-1.

Formalizing decision problems in terms of utility representations can make many issues clearer, as I hope to show in the following sections. As Rawls recognized, even deontological theories such W.D. Ross's can be faithfully represented in terms of standard cardinal utility measures.[25] But this important insight by no means licenses forgetting about what the numbers are representing, and so treating all "utility" as essentially a homogenous abstract quality subject to the same types of disputes and resolutions. If we do not keep in mind that the utility scores represent disagreements about justice, it may seem that Alf and Betty have a resource division or interest compromise problem, with the numbers indicating unequal splits of the "gains." Clearly this is not the case. Neither party has approved of the other's understanding of justice, and may in fact have severe doubts about it. Suppose Betty complains to Alf that at (R_1, R_1) she must concede 2/3, but at (R_2, R_2) he would only concede 1/3. Fair's fair, after all; and the maximum concession should be minimized.[26] But Alf might reply thus:

> "A basic reason why you concede relatively more when we both act on rule R_1 than when we both act on R_2 is that you place too much value on living according to your preferred rule of justice and too little on reconciliation; I "concede relatively less" at R_2 because I place so much more value on reconciliation than do you, even on those rules I consider to be an inferior, while you more highly evaluate living according to your "I believe we ought" judgments." You are free to do so, but I fail to see why that decision gives you a claim to additional consideration in our reconciliation — because you undervalue reconciliation!"

Alf disagrees with Betty about justice, including on the importance of reconciliation. Both the importance of acting on his "I believe we ought" judgments and of sharing a rule are factored into his utility. Note that all this is about justice *from his own perspective*, which includes a commitment to his own insights about perfect justice and the importance of living according to shared rules. That his own conception of justice leads him to take account of what rules he can share with others in no way commits him to treating his and other people's views of justice as somehow "on par" and each having a symmetric claim to the "moral gains" produced by shared rules. On reflection, the very idea of a fair distribution of moral gains between those who disagree about morality seems rather bizarre.[27]

[25] Rawls, *Political Liberalism*, p. 332n.

[26] Some have argued that common bargaining axioms are only about rationality, and have no implications concerning fairness. I believe this is wrong; motivations for the critical symmetry axiom invoke a general notion of equality. See John Thrasher, "Uniqueness and Symmetry in Bargaining Theories of Justice," *Philosophical Studies*, vol. 167 (2014): 683-699. Interestingly, in his final statement of his view Gauthier sees minimax relative concession not as a solution specifying a rational bargain, but as a standard of justice. See his "Twenty-Five On," *Ethics*, vol. 123 (July 2013): 601-24.

[27] Some argue that bargaining solutions simply predict what splits agents will agree to and are not in any way normative. The evidence for this claim is uncertain: in empirical studies people easily play one of the pure Nash equilibria in impure coordination games and, indeed, teach this to the next generation. See

4 SELF-ORGANIZATION IN MORALITY

4.1 From Constructivism to Spontaneous Orders

I have tried to stress the importance of two modes of reasoning about justice. When employing "I believe we ought" reasoning a moral agent inquires, given her standards of justice, about the extent to which rules of justice R_1 and R_2 are appropriate standards for just relations among moral agents. Given her understanding of justice, suppose that she can score every rule from 0 to 10, where 10 indicates perfect agreement with her standards and 0 indicates a rule that is entirely unacceptable as rule of justice. A critical assumption in the model to be developed is that a moral agent acknowledges acceptable approximations to her understanding of justice, as her own individual mode of deliberation judges it (scores of 1-9).[28] The reason that she would accept living according to an imperfect rule is supplied by recognition of the social character of justice: justice cannot be fully secured by unilateral action. Just social relations require participation of others, and so one's complete understanding of justice will ultimately weigh the relative importance of "I believe we ought" judgments and reconciliation.

We have seen that the social contract tradition takes seriously the social nature of justice, providing accounts of how we might collectively reason about shared rules of justice to live by. In an important sense, however, these accounts are what Hayek would call forms of "constructivism."[29] They remain versions of "I believe that [<the set of reasonable deliberators believe that> we all ought to]" judgments, in which the theorist constructs his version of the reasonable compromise among our moral views. It is not, I think, going too far to say that social contract views are "top-down" (from the philosopher to us) theories of what a "bottom-up" (what we collectively would choose) morality might look like. Once we recognize this, the question looms: what would a genuine "bottom-up" social morality look like? Such a morality, I shall argue, would be the result of each person acting on what might be called, more than a little inelegantly, her "I believe what, taking account of the justice-based choices of others, we ought to do — and that's what I ought to do" view of justice. Here there is no collective choice: the theorist seeks to inquire what would occur if each agent genuinely followed her own view of justice, taking into account her commitment to reconciliation.[30] Would agents who disagree in their (*i*) "I believe we ought" judgments of justice and (*ii*) judgments of the relative importance of reconciliation, converge on common

---

Andrew Schotter and Barry Sopher, "Social Learning and Coordination Conventions in Intergenerational Games," *Journal of Political Economy*, vol. 111 (June 2003): 498-529.

[28] See further my *The Tyranny of the Ideal*, pp. 45-9 and "The Commonwealth of Bees." We could formalize these into von Neumann – Morgenstern utility functions, but nothing critical turns on this point.

[29] F. A. Hayek, *Rules and Order* (London: Routledge, 1973), chap. 2. Cf. Rawls, *Political Liberalism*, Lecture III.

[30] I have considered in some depth the relation of this theoretical perspective to the choices of individual moral agents in "Social Morality and the Primacy of Individual Perspectives," *The Austrian Review of Economics*, DOI 10.1007/s11138-016-0358-8.

rules, or would they each go their own way? Under what conditions might free individual moral reasoning replace the collectivist constructivism of the social contract? These are the questions I shall now begin, in an admittedly simplified and tentative way, to pursue.

4.2 Justice-based Utility Functions

The following analysis is purely formal, and does not presuppose any specific substantive theory of justice or social morality. Each person is assumed to be solely interested in acting justly, as she sees it. Each is thus characterized by a justice-based utility function that represents her judgments as to the justice of states of affairs, defined in terms of what rules of justice are acted upon. Recall that — at least as I have characterized it — the aim of social contract theory is to see how free and equal moral persons who do not agree in their individual mode of reasoning about justice can share a common system of moral rules or principles. Substantive-contract attempts to secure this, we have seen (§3.1), valorize a specific conception of justice that the theorist claims "we all [could] share," but this claim itself succumbs to the problem of reasonable pluralism. Some good-willed moral agents who wish to share a system of justice with others do not accept the theorist's conclusions about what "we believe we ought to do." Consequently, I seek here a theory of shared moral life that does not advance a correct answer about justice, but instead endeavors to model how individuals who disagree about the correct answer might nevertheless converge on a common moral life.[31]

Given the analysis thus far, I assume that the justice-based utility of each person can be divided into two parts. What I shall call person *A*'s (aka Alf's) *inherent* (justice) utility of rule R_1 (denoted $\mu_A(R_1)$) represents Alf's scoring of R_1 in terms of how well it satisfies his "I believe we ought" reasoning about justice. I suppose that this ranges from 0 to 10 for all agents, with 0 representing any rule that the agent judges to be unacceptable from the perspective of justice. These utilities do not support interpersonal comparisons. Now agents who recognize the social dimension of justice also place weight on whether others share a rule. A rule R_1 where $\mu_A(R_1) = 10$, but is shared by no one, will be seen by Alf as inferior to R_2, where $\mu_A(R_2) = 9$ but is shared by all. However, again reasonable disagreement asserts itself. Good-willed competent moral agents reasonably differ on the relative importance of the two modes of moral reasoning. Some place greater weight on their "I believe we ought" reasoning (a rule's inherent utility) while others put more emphasis on the importance of reconciliation on a rule of justice. This is a critical difference that must be at the core of our analyses. To take account of moral disagreement (i.e., about inherent justice-based utility) while imposing a uniform weighting of these two modes of reasoning misses a fundamental source of our moral differences.[32] We thus suppose that each person has a

[31] On this conception of political philosophy, see Fred D'Agostino, "How Can We Do Political Philosophy?" *Cosmos + Taxis* (forthcoming).

[32] In this sense the present analysis presses beyond that in *The Order of Public Reason*.

weighting function that takes account of how many others act on a rule — how widely it is shared — and the importance to her of that degree of sharing. This weighting function for person *B* (aka Betty) will be denoted as $w_{B(n)}(R_1)$, which is the weight that Betty gives to Rule R_1 when *n* others act on it. I suppose weights vary between 0 and 1.

There are an infinite number of weighting systems. Figure 4 presents the ones on which we shall focus, here with *n* = 101

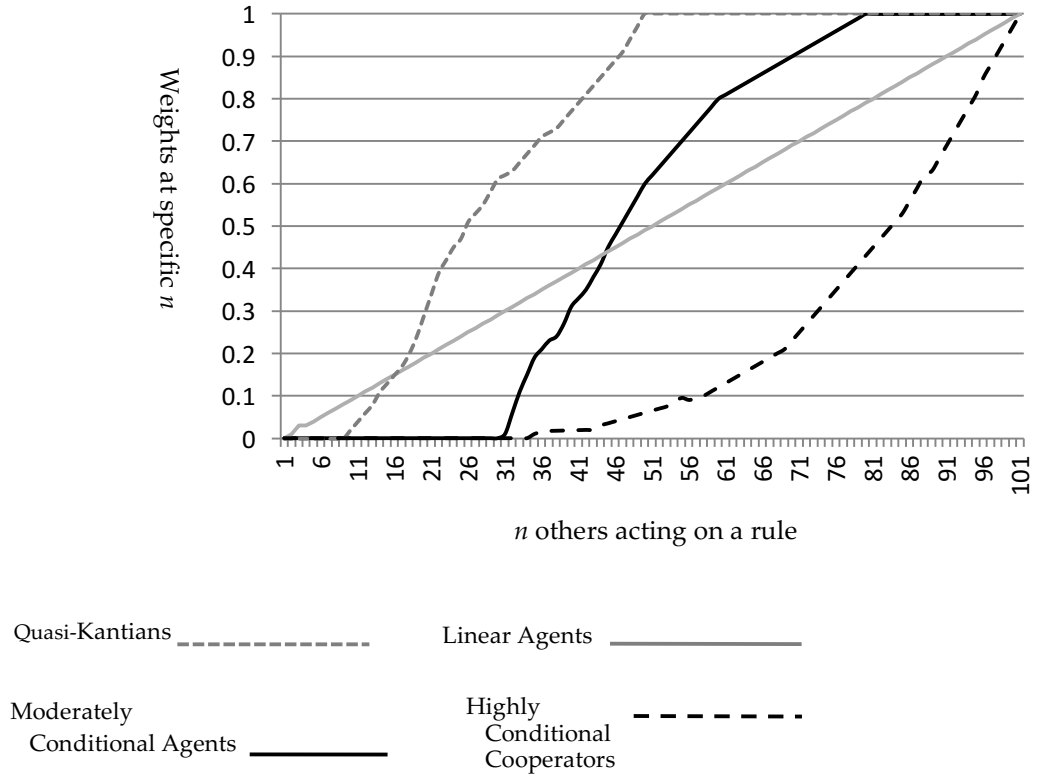

Figure 4: Four Weighting (Reconciliation) Types

All four types acknowledge that reconciliation is part of their view of justice, so appreciating the social dimension of justice. "Quasi-Kantian" agents recognize some value of reconciliation: they give no weight to a rule that is not practiced by 10% of the population, but by 50% they give a rule a maximal weighting of 1. Moderately Conditional Agents have similar shape to their weighting function but are more typical of Humean conditional cooperators: until a significant share of the population acts on a rule they are not willing to act, and so give it a 0 weight.[33] A rule must have 30% uptake before they give it any weight, and reaches a maximal weight at 80%. Both Quasi-Kantians and Moderately

[33] Cristina Bicchieri models conditional cooperators as having a certain threshold, often requiring that "most" others share a rule before they will act on it. Our types do not have abrupt thresholds, but Moderately Conditional Cooperators have a significant threshold of about 30%. See her *The Grammar of Society* (Cambridge: Cambridge University Press, 2006), pp. 11ff.

Conditional Agents, we might say, seek a moral community but not a maximally large one. Linear Agents have, unsurprisingly enough, a linear weighting function: the more share the merrier, but as long as someone acts on their rule, they give it some positive weight. Lastly, Highly Conditional Cooperators are resolute in stressing the importance of reconciliation, and only weight rules highly when the large majority has already joined in. These Highly Conditional Cooperators are, as it were, willing to play the justice game only if most others do. Highly Conditional Cooperators give no weight to a rule unless about a third of the population follows it, and give very little weight to any rule practiced by less than 60%. They do not give really high weights until approximately 90% practice it. They are thus *highly* conditional moral agents. Highly Conditional Cooperators are in some ways the mirror image of our Quasi-Kantians, perhaps died-in-the-wool Humeans.

In this analysis, then, an agent is concerned with both his own evaluations of the inherent justice of a rule (i.e., his "I believe we ought" conclusions) and reconciliation with the judgments of others ("we believe we ought"), and will ultimately make his decision based on his own view of the inherent justice of the rule given his evaluative standards and the weighted number of others who are acting on the rule. Letting U_A be Alf's total justice-based utility of acting on rule R_i, $\mu_A(R_i)$ the inherent justice-based utility of R_i given Alf's evaluative standards, w_A his social weighting and n the number of people acting on R_i, we get:

$$\text{EQ. 1} \qquad U_A(R_i) = \mu_A(R_i) \times w_{A(n)}(R_i)$$

If Alf is confronted by two rules, he will act on that which maximizes U_A. So Alf acts on R_1 rather than R_2 only if $U_A(R_1) \geq U_A(R_2)$.[34]

4.3 Agent-based Modeling and Parametric Rationality

Gauthier usefully distinguishes two types of rational choice contexts. In "*parametric* choice…the actor takes his behavior to be the sole variable in a fixed environment. In parametric choice the actor regards himself as the sole center of action. Interaction action involves *strategic* choice, in which the actor takes his behavior to be but one variable among others, so that his choices must be responsive to his expectations of others' choices, while their choices are similarly responsive to their expectations."[35] As with most theoretical distinctions, this one is perhaps not quite so crystal clear as it first seems, but it highlights an important difference in rational choice. In the Go Your Own Way (Figure 1) and Reconciliation Games (Figures 2 and 3), Alf's action was dependent on what he thought Betty was going to do and what she chose to do depended on what she thought he was going to do. Indeed, ultimately Alf's choices depended on considerations such as what he thought she thought he thought she was going to do, and so on. Thus the crucial

[34] And he *will* act on R_1 if $U_A(R_1) > U_A(R_2)$.

[35] Gauthier, *Morals by Agreement*, p. 21. Emphasis in original.

importance of common knowledge assumptions in game theory, whereas in situations of parametric choice a person takes the actions of others as a given: what is her best course of action *given* the actions of others, over which she has no influence? In contrast to strategic situations she supposes that her choice will not affect the choices of others; their choices are taken as a parameter — a given or constraint — in her decision. Thus, as Gauthier recognizes, rational choice of a consumer in a large market is quintessentially parametric — she simply adjusts her action to the given prices; compare a traditional shopping bazaar in which she is a price-taker *and* setter, and so is engaging is strategic interaction. In parametric contexts an individual still may seek an equilibrium result: i.e., one in which given the context in which she finds herself she cannot unilaterally increase her utility by a change in her choice.

As I said, as with most distinctions this one becomes less clear as we inspect it more closely. In a large market at any given time each consumer is a price taker, not a price setter, but over a series of periods in the market each consumer's parametric choices affect the price, and are minute influences in setting the price in subsequent periods. So it is an exaggeration to say that one's choices have no effect on others' choices.[36] The critical point is that at each choice node, one takes the actions of others as an exogenous variable that is simply a parameter in one's maximizing decision. It is also something of an exaggeration to say that in parametric choice expectations about others do not matter, as such expectations may be a critical parameter. If one is on one's way to the airport and wishes to minimize travel time by choosing the fastest route, one confronts a parametric choice, but of course one's empirical beliefs about which road is presently congested (one's empirical expectations about how much traffic one will encounter on each route) is critical. Parametric choice requires empirical beliefs about what others are doing, so that one can efficiently respond.

Moral and political philosophers tend to underappreciate the insights of parametric analysis. Indeed, even Gauthier, who has a much deeper appreciation of it than most, holds that morality only enters when strategic choice arises.[37] Strategic choices are often small-number interactions, and so "ppe-inclined" philosophers, often wedded to the strategic outlook, have repeatedly analyzed morality in terms of small-number (very often dyadic) interactions. We know, though, that a social system of justice is typically a large-numbers phenomenon, in which many individuals interact, each adjusting her action to what she sees as the current social parameters. As generations proceed, the aggregation of parametric choices changes the social parameters (as with prices in large markets), which in turn

[36] This is a familiar point in agent-based models of adaptive system: "…as agents adjust to their experiences by revising their strategies, they are constantly changing the context in which other agents are trying to adapt." Robert Axelrod and Michael D. Cohen, *Harnessing Complexity* (New York: Basic Books, 2000), p. 8. In evolutionary game theory replicator dynamics can make it look as if players are strategically responding to each other. In my view they best model evolutionary adaptive systems, which are my concern here.

[37] Gauthier, *Morals by Agreement*, p. 21. But cf. pp. 170-1.

changes what is parametrically rational. Here parametric models can be most enlightening, which suppose that each person has a utility function (a representation of a preference ordering)[38] and beliefs about the state of the world at time *t* (which typically include what others are now doing), and each acts to maximize her utility. In some agent-based parametric models the system will stabilize in the sense that, at some iteration *i*, further states of the system confront all the members with exactly the same parameters such that no one henceforth changes their choices (leaving aside preference change, errors in beliefs, and factors exogenous to the system). In other systems there can be endless adjustments; they will never reach system-wide equilibrium.

All the following analyses are agent-based models of the former sort. Each agent has a simple binary option set of acting on either R_1 or R_2, which are assumed to be alternative rules of justice over some area of social life, say property rules, promising rules, privacy rules, etc.[39] It is also assumed that whether a person has acted on R_1 or R_2 in the last period is reliable public information; in period *i*, each has a correct first-order belief about what rules of justice others endorsed and acted upon in period $i - 1$. Our aim is to get some preliminary insights when free moral agents, each with her own distinctive justice-based utility function and fully committed to acting on it, are apt to converge on a shared rule of justice, and under what conditions they are less likely to.

5 SOME DYNAMICS OF SELF-ORGANIZATION

5.1 Model I: Fully Random

We start with a population of 101 agents, with $\mu(R_1)$ and $\mu(R_2)$ scores between 1 and 10 that were randomly assigned. Thus all agents hold each rule is an acceptable approximation of justice, if only barely (a score of 1). Agents were randomly assigned to one of our four weighting types.[40] In the first period each individual simply acts on her "I believe we ought" judgment, maximizing her evaluative utility (μ). In ties [i.e., $U(R_1) = U(R_2)$]) an individual acts on R_1; perhaps R_1 is the simpler rule, and so individuals choose it in a tie. Our empirical updating rule is simple, if somewhat dumb: at each period an agent calculates whether in the previous period she would have achieved more utility if she had acted on the alternative rule; if she would have she switches in this period.

As Figure 5 shows, Rule 2, randomly favored by 51 agents compared to R_1's 50, went to fixation after five periods. The explanation is that valuing reconciliation with others is an increasing returns dynamic. W. Brian Arthur famously showed that when a good is

[38] Again, it is important to always keep in mind that a preference is simply a binary relation according to which one state of affairs (action, etc.) is better than another; it is not a reason or explanation as to *why* one is better than another. To say that Alf holds that "α is preferred to β" because that "is his preference" is a tautology, not an explanation. A utility function is a mathematical representation of a consistent preference structure.

[39] There are some difficulties in further formalizing this idea. See *The Order of Public Reason*. pp. 267ff.

[40] For precise weightings, see the Appendix.

characterized by increasing returns — when the more others use it the more valuable it is to any single user — the possibility of cascades arises.[41] For all four of our weighting functions, over some range of *n*, the more people act on R_i the greater weight the agent gives it — the more value she puts on it. Each of our agents is seeking to balance devotion to her "I believe we ought" judgments with some valuing of reconciliation. Within some range, the larger the number of others who share a rule, the greater its reconciliation benefits. While Alf may not switch from R_1, supported by his individual mode of moral judgment when 50% of the population is acting on it, as the number shrinks (that is, as others come to endorse R_2), the reconciliation benefits of acting on R_2 are apt to eventually be so great that he will abandon R_1. Note that this is not because he had changed his relative evaluation of the importance of the two modes of reasoning, but because the justice-value of reconciliation becomes so large as more and more others endorse R_2.

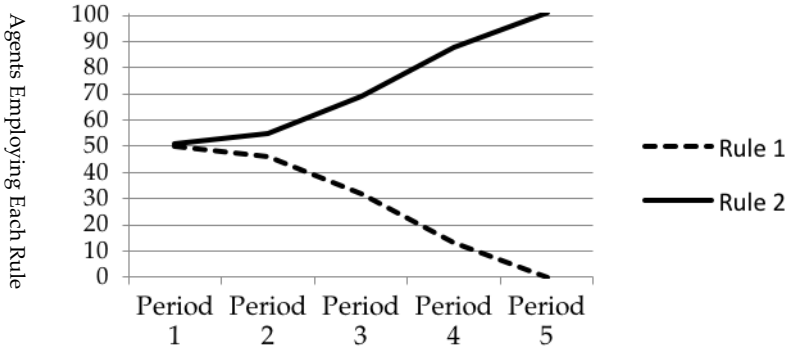

Figure 5: Convergence in Model I, All Types

A cascade occurred in Model I. The first agents to switch from R_1 to R_2 were those that, while their inherent utility deems R_1 slightly more just that R_2, place a sufficiently high value on reconciliation with others such that the greater benefits of reconciliation on R_2 outweighed the slight inherent justice advantage of R_1. Four agents made this reevaluation, but this was enough to get a cascade going. After they switched to R_2, others, perhaps who thought the inherent justice advantage of R_1 over R_2 was slightly greater than our first group (and so did not switch in period 2), come to the conclusion that, given the slightly greater number of people following R_2 in period 2, the reconciliation benefits of R_2 now outweighed R_1's inherent justice advantage, and so they changed in period 3 (14 agents did this). And we can see that in period 4, many of those who thought R_1 was considerably

[41] W. Brian Arthur, *Increasing Returns and Path Dependence in the Economy* (Ann Arbor, MI: University of Michigan Press, 1994). See also *The Order of Public Reason*, pp. 389-400.

superior to R_2 now came to the conclusion that an insufficient number were acting on R_1, and so it provided insufficient reconciliation (23 agents). By period 5 all R_1 supporters decided to endorse R_2.

It might be wondered whether any specific weighting type was crucial in producing convergence, but as Figure 6 shows, under this same population, any three of the weighting systems (again, randomly assigned) resulted in fixation on R_2, giving some reason to believe that the convergence dynamic is not driven by specific types. However, we do see that combinations of types certainly have an effect; omitting the Highly Conditional Cooperators, for example, slowed down convergence.

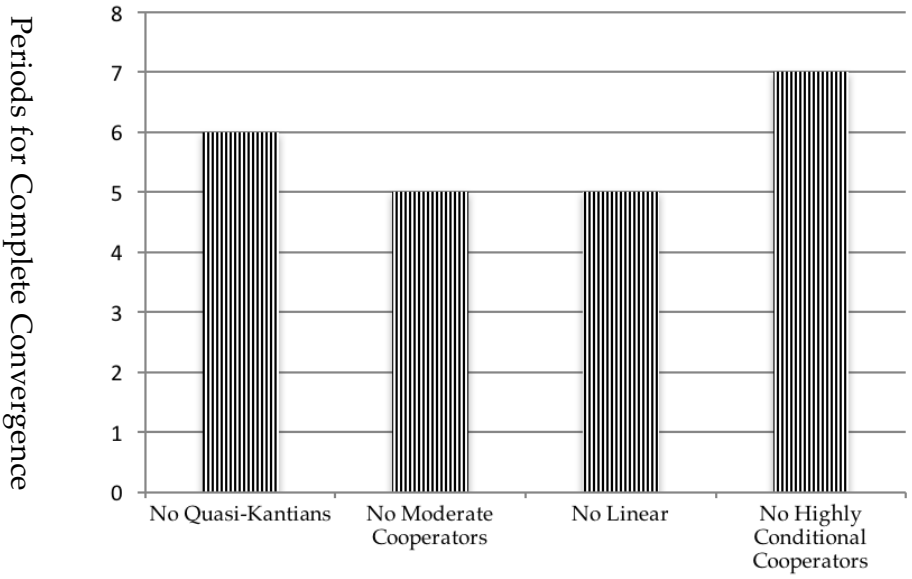

Figure 6: Periods for Convergence in Model I, Three Types

Finally, it might be wondered what occurs if some of the agents have no interest in reconciliation: as they understand "pure" or "maximal" integrity, it requires always acting on their "I believe we ought" judgments. Of course given a large enough contingent of such agents, if they disagree there will be no convergence; but what if, say, of 10% of the agents are of this sort? To see if such agents easily block convergence, 10% of the R_1 favoring agents in our population were replaced by such "maximal integrity agents." We might hypothesize that these agents would either tilt the convergence dynamic to R_1, or at least slow down the process converging on R_2. In fact, they had no effect on the behavior of others: convergence of the rest of the population (96 agents) on R_2 occurred in the same number of periods (five), with the only difference being that these resolutely non-

reconciliation agents formed their own, five-member R_1 network.

5.2 Model II: Moderate Polarity

I commenced with a fully random model to explore the core dynamics under conditions where the population was very closely divided, and to better see some of the effects of the different weighting systems. It certainly is clear that the dynamic does not depend on one rule having an initial overwhelming advantage (51/50 was enough). Different weightings have different thresholds and values, which help induce convergence dynamics. However, fully random distributions of inherent utility and weighting functions are hospitable to cascades, since they tend to ensure that there will be continuity of degrees of $U(R_1) - U(R_2)$ differences, such that whenever some agents switch, this will decisively affect the choice of the next agents "in line," who then switch in the next period, and so on. The question is the extent to which a convergence dynamic applies in non-random populations with discontinuities. An especially difficult case is a polarized population, divided into two mutually exclusive groups, one subgroup thinking highly of one option and scoring the other low, with the other subgroup doing the opposite.

To explore this possibility our group of 101 agents was divided into two "Hi-Lo" groups, one of which scored R_1 between 10 and 6, and R_2 between 4 and 1 [thus $\mu(R_1)$ are all "high," while $\mu(R_2)$ are all "low"]; the other group was assigned the opposite "Hi-Lo" inherent utilities. Each polarized group had approximately an equal division of all four types of agents. The suspicion that polarization makes convergence more difficult was confirmed; very closely split [52 agents $\mu(R_1)$ High; 49 agents $\mu(R_2)$ High] polarized populations did not display tendency to converge. Somewhat surprisingly, perhaps, convergence on R_1 did occur within four periods at the close but not finely balanced [56 agents $\mu(R_1)$ High; 45 agents $\mu(R_2)$ High]. As Figure 7 shows, it was found at this 56/45 division complete convergence occurred with any three types; again the omission of Highly Conditional slowed down the process (taking 8 periods), while omitting the Moderately Conditional Cooperators slowed convergence to 7 periods. Conditional Cooperators, of course, generally put greater weight on reconciliation, and so assist in overcoming polarity. Indeed, a Hi-Lo polarized population of all Highly Conditional Cooperators with an initial 53/48 advantage for R_1 — a pretty evenly split polarized population — secured complete convergence on R_1 in four periods. Quasi-Kantians, on the other hand, tend to reinforce the split; they have their maximal weightings at around 50% of the population, and so tend to reinforce polarity when the groups are about the same size. Quasi-Kantians who find themselves in subgroups who agree with them do not easily switch rules. Perhaps the truly striking thing is that even they can be induced to leave their Hi-Lo groups and converge on a common rule, and do so when any two of the other weighting types are well represented. If Hi-Lo polarity is not too finely balanced, then, it can be overcome in a diverse population: the diversity of weighting types typically speeds up the process, inducing sufficient continuity in the populations' $U(R_1) - U(R_2)$ differences even with the population is

characterized by polarized (thus discontinuous) $\mu(R_1) - \mu(R_2)$ differences.

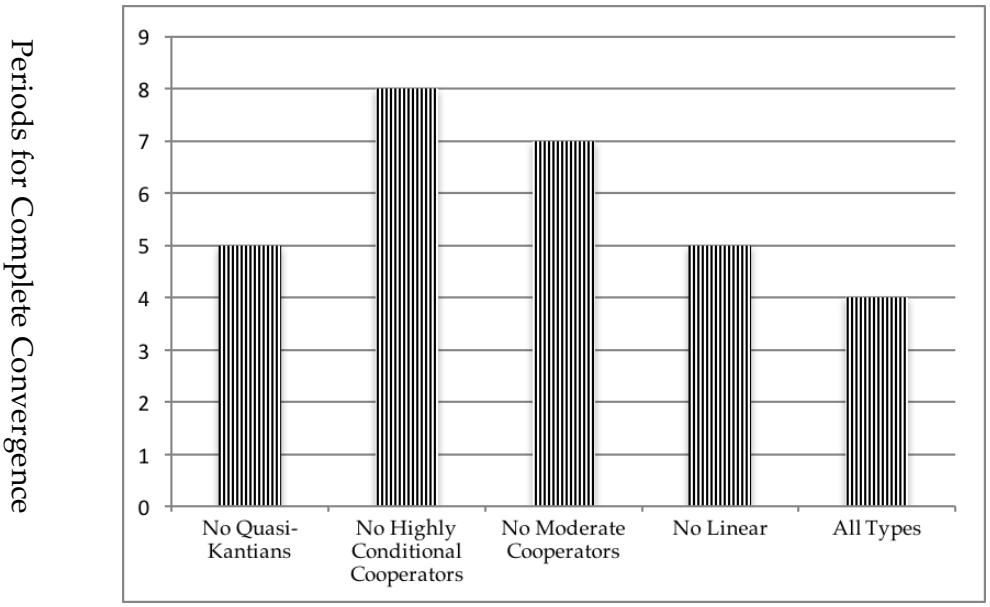

Figure 7: Periods for Complete Convergence in Model II

While it is not surprising that a population composed entirely of those who put great value on reconciliation (The Highly Conditional Cooperators, for example) can overcome a polarization in a population's "I believe we ought" judgments, it is, I think, worthy of emphasis that as we add diversity of weighting types, polarization can be more easily overcome than in many more homogenous populations: in Figure 7 convergence was quickest when all four types were present. This could point the way to good news, for only not only do we disagree about justice, but for the last hundred and fifty years western societies have been significantly polarized between "right" and "left" justice, with the last forty adding a number of other groups (e.g., feminists, environmentalists) who also tend to "Hi-Lo" judgments. Rather than reasonable pluralism we should, perhaps, be thinking of moderate reasonable polarity. The Polarity Model gives us some reason to suppose that these sharp inherent justice differences can be significantly moderated by a diversity of weighting types. It is not necessary that we all highly value reconciliation to overcome polarization. An uptake of this idea would constitute a fundamental change in the orientation of the public reason project, which has thus far supposed that diversity is the problem, and commonality the sole route to sharing. Here, we see the possibility that one type of diversity can counteract the centrifugal tendencies of another. So far from heterogeneity always be an impediment to convergence on a shared rule of justice, some configurations of diversity can help secure agreement. The issue is not "do we agree enough to live together?" but "does the overall pattern of homogeneity and heterogeneity

induce convergence on common ways of living together?"

5.3 Model III: Differential Reference Groups

A simplification in the models thus far discussed is that each person takes the entire group as her reference group: in each period her decision about reconciliation is based on what the entire group has done. But often people are concerned with narrower reference groups.[42] Alf might seek to reconcile his view of justice with his traditional cultural group while Betty's reconciles with those in her urban and work environments. In previous work I have supposed that we seek a practice of moral accountability based on shared moral rules with the widest feasible set of other moral agents.[43] But in many contexts people might be committed to a practice of accountability — and so the shared understanding of the rules of justice on which it depends — only with those with whom they regularly interact, while others may be interested in a practice of accountability based on shared rules with some other group. In this case the different elements of the population would have different reference groups — different groups of people with whom they seek to reconcile. Can there be shared rules by moderately polarized groups under such circumstances?

To take some first steps in understanding the effects of different reference groups on convergence, let us analyze a somewhat challenging case: the population is not only split into different reference groups, but some of the reference groups display opposite Hi-Lo polarity. In group B, 3/5 of the population has Hi-Lo bias in "I believe we ought" judgments in favor of R_1 (the other 2/5 of the group has Hi-Lo inherent justice utilities in favor of R_2), while the C group has just the opposite Hi-Lo divisions. Here, we might think, convergence within each group will occur, but not between them: our polarized population models in section 5.2 indicate that with such splits we should expect convergence on the most popular rule in each group. And of course that is what would normally happen if these are entirely unrelated reference groups, for then we simply have two independent populations. The interesting case concerns populations with overlapping reference groups. In the Differential Reference Group Model a population of 150 agents is divided into three main groups, with two of the groups having subgroups. They are:

> Group A (50 agents): split population, not Hi-Lo (an agent may have any combination of $\mu(R_1)$ and $\mu(R_2)$ between 1 and 10; inherent justice utilities are randomly assigned).
>
> Group B2 (25 agents): approx. 3/5 Hi-Lo favoring R_1; 2/5 Hi-Lo favoring R_2.
>
> Group B1 (25 agents): approx. 3/5 Hi-Lo favoring R_1; 2/5 Hi-Lo favoring R_2.
>
> Group C1 (25 agents): approx. 3/5 Hi-Lo favoring R_2; 2/5 Hi-Lo favoring R_1.
>
> Group C2 (25 agents): approx. 3/5 Hi-Lo favoring R_2; 2/5 Hi-Lo favoring R_1.

[42] Cristina Bicchieri extensively examines the role of reference groups in her *Norms in The Wild* (Oxford: Oxford University Press, 2016). I made some preliminary remarks on this point in *The Tyranny of the Ideal*, pp. 184-7.

[43] See *The Order of Public Reason*, pp. 279-83.

Note that both subgroups in B have identical evaluative utility distributions, as do both subgroups in C. The difference is their reference groups, as indicated by Figure 8.

Group	*Reference Group*
B2: R_1 Hi-Lo Biased, Parochial (25 agents)	**B1, B2**
B1: R_1 Hi-Lo Biased, Involved (25 agents)	**B1, B2, A**
A: Random Group (50 agents)	**B1, A, C1**
C1: R_2 Hi-Lo Biased, Involved (25 agents)	**C1, C2, A**
C2: R_2 Hi-Lo Biased, Parochial	**C1, C2**

Figure 8: Differential Reference Groups

Group B2, then is "Parochial R_1 Hi-Lo Biased" insofar as they update only in relation to the choices of Group B, the Hi-Lo R_1 biased group, in the previous period. This means that Group B2 (*i*) has a reference group of 50 agents and (*ii*) their entire reference group has a 3/5 Hi-Lo bias toward R_1. B1's reference group is 100 agents, encompassing all of Groups A and B. I call this an "Involved Hi-Lo Biased" group as it is concerned with reconciliation both with its entire R_1 Hi-Lo biased group as well as the "wider" world of A. Group A, the Random group, has a reference group of 100 agents, including all of Group A itself, as well has half of both Hi-Lo biased groups (B1, which is biased toward R_1, and C1, which is biased towards R_2). Group C is the mirror image of Group B. Note that we have two parochial subgroups, B2 and C2, whose reference networks are restricted to those who share their inherent justice Hi-Lo distributions.[44]

As in other models, each agent simply acts on her judgment of the inherent justice of R_1 or R_2 in the first period. However, because we have multiple reference groups that employ different updating calculations, an order of updating was applied in all periods after 1; first Group A updated and acted, then B1, B2, C1 and finally C2. Thus each period has five mini-periods; when a group updates it considers the last move made by others in its reference group. This is by no means an entirely innocent stipulation: as Arthur pointed out, in closely split populations those who move earlier can have significant effects on the outcome of convergence.[45] As some counterweight to the polarized groups B and C, the random group, A, was thus given the first move in each period, giving random factors some

[44] In these two subgroups with reference groups of 50 (B2, C2), all weighting systems were normalized so that maximum $n = 50$.

[45] Arthur, *Increasing Returns and Path Dependence in the Economy*, chap. 5.

advantage over the effects of Hi-Lo polarity.

In the basic simulation, Group A had a 54% to 46% tilt towards R_1; Group B was Hi-Lo polarized 60/40% towards R_1, while Group C was Hi-Lo polarized 62/38% towards R_2 (Group C2 had a 64% having Hi-Lo bias towards R_2). This resulted in the entire population group of 150 favoring R_1 51% to 49%, a very closely split population with a good deal of polarity. Agent types were equally distributed among all five groups. As Figure 9 shows, convergence on R_1 occurred in ten periods.

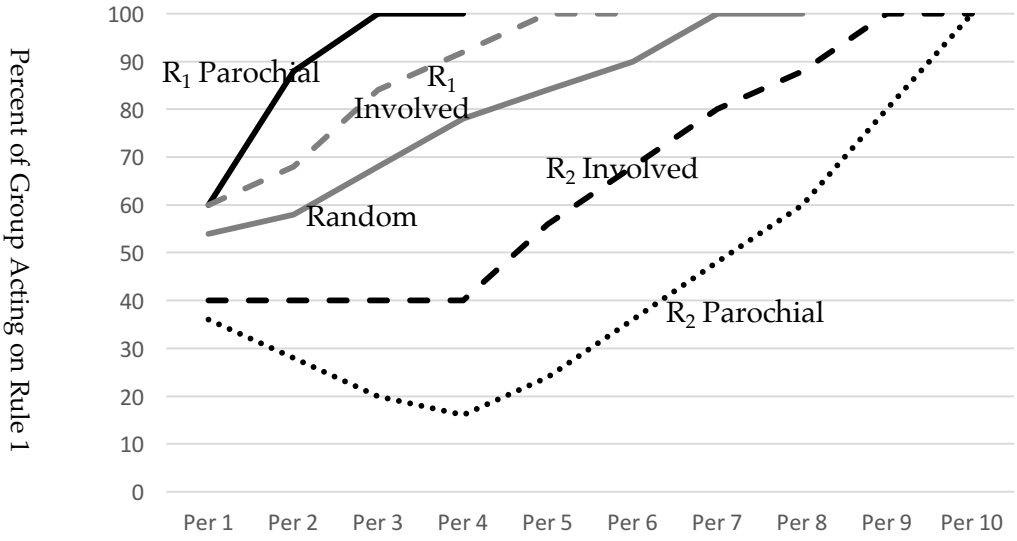

Figure 9: Convergence: Five Differential Reference Groups with Hi-Lo Polarity in Four Groups

Interestingly, the Parochial R_2 Biased group (64% Hi-Lo bias towards R_2) and whose entire reference group also had a 60% bias toward R_2, began, as we would expect from our analysis in the last section, by moving *towards* R_2: at one point being 84% R_2 followers. The Involved R_2 Biased group (C1) was, at it were, initially pulled in two directions: some of their reference group (A) was moving towards convergence on R_1, while C2 was moving toward R_2. For several periods, then, the Involved R_2 Biased group remained unchanged, until the movement in Group A was strong enough to pull them towards R_1. And, in turn, that eventually pulled C2, the Parochial R_2 Biased group, in their wake, ending up with 100% R_1 convergence. The last to switch to R_1 were, as would be expected, Quasi-Kantians favoring R_2 in the Parochial R_2 Biased group (and one Linear Agent).

Diversity of types is important, though not always necessary — if we have the right sort of type. In a simulation with the same distribution of inherent utilities as above, a homogeneous population of Linear Agents, for example, failed to converge on a rule in *any* of our five groups; the same occurred in a pure population of Quasi-Kantians. In a homogeneous population of Highly Conditional Cooperators, however, convergence was

quickly achieved, in five periods. This should not be surprising, since Highly Conditional Cooperators give great weight to high convergence. As I said, they are died-in-the wool Humeans. Nevertheless, the important and surprising lesson from our simpler models is confirmed: sometimes adding *more diversity* makes agreement more likely.

In another Model III simulation with all four types, the Random group's (A's) inherent utilities were randomly reassigned, resulting in 66% of A favoring R_1. Not surprisingly, given this strong initial tilt to R_1, full convergence was achieved in the population of 101 agents very quickly — in four periods. More interesting is what occurred when not only inherent utilities, but *agent types* were randomly distributed in the population (Figure 10).

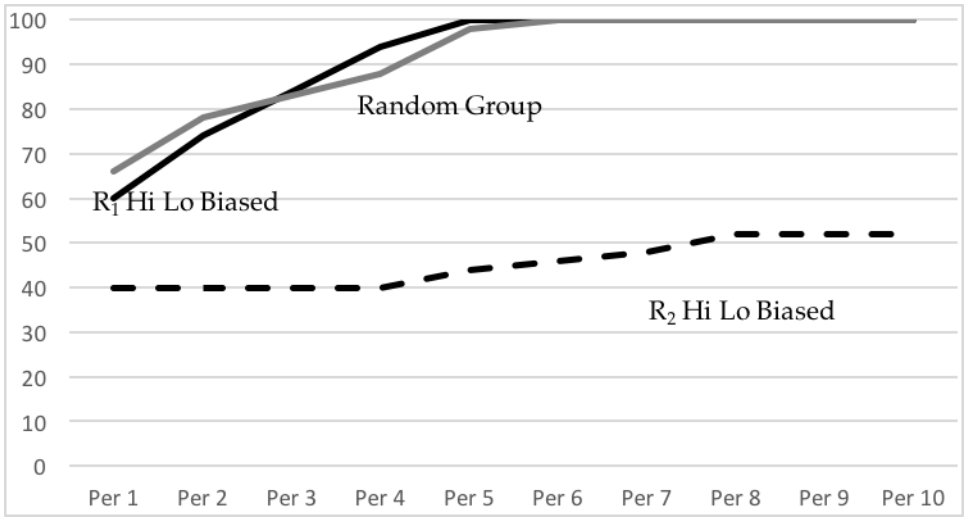

Figure 10: Random Distribution of Types (Linear Agent and Quasi-Kantians Halt Convergence)

Here full convergence on R_1 was not achieved: 24 (of the 50) members of Group C (including all the Hi-Lo biased members of the Parochial R_2 Biased group) maintained a small R_2 network, while the rest of the population (126 agents) moved to R_1. In Group C1 (the Involved R_2 Biased group), those with Hi-Lo evaluative utilities biased in favor of R_2 were almost all Linear Agents (with three Quasi-Kantians). As a result, those Hi-Lo biased in favor of R_2 in C1 were not very sensitive to movement in Group A to R_1, which in turn insulated C2 (the Parochial R_2 group) from the movement in the Random group, A. Linear agents engage in such gradual updating that few could overcome their Hi-Lo bias (in inherent justice judgments) in favor of R_2, even though there was some movement within the Involved R_2 Biased group (C1) to R_1. Recall that in a homogenous population of Linear Agents none of the five groups achieved convergence. Again, we see how a diversity of types can generate agreement: in the R_2 Biased group there was not enough diversity of weighting types to overcome its Ho-Lo bias.

5.4 Modeling Moral Choice

All three of these models are rather basic analyses of *perfectly moral and rational* agents, providing an initial exploration of some dynamics of rational choice that lead a population with somewhat stark moral disagreements to reconcile on a shared moral rule. Like social contract theories such as Rawls's, the point of these models is to understand rational moral persons and their choices, not to make predictions about actual systems of interactions.[46] The models seek to capture only motivations based on an agent's devotion to her "I believe we ought" judgement and the weight she puts on reconciliation in her view of justice. I have tried to show here that under some conditions characterized by deep diversity perfectly moral agents would be able to organize themselves into freely-endorsed moral systems. At least from a purely normative point of view, central moral controllers (such as social contract theorists) are not necessary to secure a result that all would endorse, given their differing views of justice. A number of parameters are relevant to these models: agents' information about others, the depth and extent of their moral disputes, weighting functions, sizes of, and links among, reference groups and so on. These models investigated some of these in a preliminary way.

6 MORAL FREEDOM AND UNITY IN DIVERSE, COMPLEX SOCIETIES

Contemporary moral theory and social philosophy divides into two opposing lines of thought or, we might say, research projects. The traditional, still dominant, project carries on with articulations of the "I believe we ought" view. These accounts are often sophisticated and admirable exercises in philosophical reasoning, building on the fundamental intuition that the best moral conclusion for one is the best for all. This research project has great difficulty in even making sense of the idea of moral diversity.[47] To be sure, there is moral disagreement and conflict — some pig-headed and ill-informed and some, perhaps, more reasonable — but the study of ethical life need be no more focused on diversity than is physics.[48] The second line of inquiry seeks, in disparate ways, to make sense of the idea of fundamental moral difference. To Isaiah Berlin, the Romantic philosophers of the seventeenth and eighteenth centuries have "permanently shaken the faith in universal, objective truth, in matters of conduct" by showing that "ends recognized as fully human are at the same time ultimate and mutually incompatible."[49] Berlin

[46] Which is not to say that the self-organizing systems approach here is not relevant to the study of actual systems: it connects up with Bicchieri's project on the empirical study of actual normative self-organized networks. See her *Norms in the Wild*.

[47] I have greatly benefitted from discussing the issues raised in this section with Piper Bringhurst.

[48] This, of course, can be a two-edged comparison. See Thomas Kuhn, *The Essential Tension* (Chicago: University of Chicago Press, 1977) and D'Agostino, *Naturalizing Epistemology* (London: Pelgrave, 2010).

[49] Berlin, "The Apotheosis of the Romantic Will" in *The Crooked Timber of Humanity: Chapters in the History of Ideas,* edited by Henry Hardy (Princeton: Princeton University Press, 1990): 207-37 at p. 237. See also Berlin, *The Roots of Romanticism* (Princeton: Princeton University Press, 1999), pp. 34ff; Bernard Williams, "Conflict of Values" in his *Moral Luck* (Cambridge: Cambridge University Press, 1981), pp. 71-82.

repeatedly insisted that the Romantics taught us that there are many values, and that they are incommensurable; "the whole notion of plurality, of inexhaustibility, of the imperfection of all human answers and arrangements, the notion that there is no single answer which claims to be perfect and true ... all this we owe to the romantics."[50] I have tried to show here that this second line of inquiry — which somehow takes the diversity of moral conclusions as a basic datum of ethical inquiry — is fundamental to the social contract tradition. Once such diversity is understood not as moral reasoning gone awry, but as the crux of free human moral reasoning, moral diversity in some guise becomes the core of moral theory and social philosophy.

Yet the social contract — and this most definitely includes Rawls — never really advanced beyond what we might call "plans to manage difference."[51] Moral difference is seen as the fundamental problem for moral theory, but the aim is to plan for mediation via a social contract that rises above difference to show an underlying homogeneity. Certainly mediation and reconciliation are fundamental concerns of moral theory, for as soon as we cannot suppose that good moral reasoning alone shows us the path to a cooperative social life, we need to find new paths, which produce some unity in moral expectations and understandings *out of diversity*. F. A. Hayek stressed throughout his career that rational constructivism and planning is not usually a viable way to cope with heterogeneity. Central planners, be they economists or moral philosophers, do not have access to the necessary information about diversity: they can only cope with it by limiting admissible diversity, relying on "normalizing" assumptions about agents.[52] I have tried to take some small steps here in theorizing how we might think of moral theory without that form of central planning practiced by social contract theorists. The guiding idea is to model morally autonomous diverse agents making choices in the context of each other's choices, seeing what dynamics lead to a shared rule that all endorse, and when different groups will go their own way. The motto of this project is that morality is best understood as a bottom-up affair. "The moral law is not imposed from above or derived from well-reasoned principles" but arises from the values of individuals and their distinctive searches for integrity and reconciliation in their social-moral lives.[53]

[50] Berlin, *The Roots of Romanticism*, p. 146. Rawls was sufficiently schooled in the first line of inquiry that he draws back from Berlin's talk of "competing truths" in morality, seeking to put questions of truth aside and focus on the concept of the reasonable. See Rawls, "The Independence of Moral Theory."

[51] An important exception to this broad claim is Muldoon's *Social Contract Theory for a Diverse World* (New York; Routledge, 2016) — though his is a rather unusual contract, sharing much with a self-organizing analysis.

[52] See my *Tyranny of the Ideal*, esp. chap. 4.

[53] Frans de Waal, *The Bonobo and the Atheist* (New York: W.W. Norton, 2013), p. 228. This essay's epigraph was from page 23.

N =	MC	LA	QK	HCC
1	0	0	0	0
2	0	0.01	0	0
3	0	0.03	0	0
4	0	0.03	0	0
5	0	0.04	0	0
6	0	0.05	0	0
7	0	0.06	0	0
8	0	0.07	0	0
9	0	0.08	0	0
10	0	0.09	0.02	0
11	0	0.1	0.04	0
12	0	0.11	0.06	0
13	0	0.12	0.08	0
14	0	0.13	0.11	0
15	0	0.14	0.13	0
16	0	0.15	0.15	0
17	0	0.16	0.17	0
18	0	0.17	0.2	0
19	0	0.18	0.24	0
20	0	0.19	0.29	0
21	0	0.2	0.34	0
22	0	0.21	0.39	0
23	0	0.22	0.42	0
24	0	0.23	0.45	0
25	0	0.24	0.47	0
26	0	0.25	0.51	0
27	0	0.26	0.53	0
28	0	0.27	0.55	0
29	0	0.28	0.58	0
30	0	0.29	0.61	0
31	0.01	0.3	0.62	0
32	0.06	0.31	0.63	0
33	0.11	0.32	0.65	0
34	0.15	0.33	0.67	0
35	0.19	0.34	0.69	0.01
36	0.21	0.35	0.71	0.015
37	0.23	0.36	0.72	0.017
38	0.24	0.37	0.73	0.0175
39	0.27	0.38	0.75	0.018
40	0.31	0.39	0.77	0.0185
41	0.33	0.4	0.79	0.019
42	0.35	0.41	0.81	0.0195
43	0.38	0.42	0.83	0.02
44	0.41	0.43	0.85	0.03
45	0.45	0.44	0.87	0.035
46	0.48	0.45	0.89	0.04
47	0.51	0.46	0.91	0.045
48	0.54	0.47	0.94	0.05
49	0.57	0.48	0.97	0.055
50	0.6	0.49	1	0.06
51	0.62	0.5	1	0.065
52	0.64	0.51	1	0.07
53	0.66	0.52	1	0.075
54	0.68	0.53	1	0.085
55	0.7	0.54	1	0.095
56	0.72	0.55	1	0.09
57	0.74	0.56	1	0.095
58	0.76	0.57	1	0.1
59	0.78	0.58	1	0.11
60	0.8	0.59	1	0.12
61	0.81	0.6	1	0.13
62	0.82	0.61	1	0.14
63	0.83	0.62	1	0.15
64	0.84	0.63	1	0.16
65	0.85	0.64	1	0.17
66	0.86	0.65	1	0.18
67	0.87	0.66	1	0.19
68	0.88	0.67	1	0.2
69	0.89	0.68	1	0.21
70	0.9	0.69	1	0.23
71	0.91	0.7	1	0.25
72	0.92	0.71	1	0.27
73	0.93	0.72	1	0.29
74	0.94	0.73	1	0.31
75	0.95	0.74	1	0.33
76	0.96	0.75	1	0.35
77	0.97	0.76	1	0.37
78	0.98	0.77	1	0.39
79	0.99	0.78	1	0.41
80	1	0.79	1	0.43
81	1	0.8	1	0.45
82	1	0.81	1	0.47
83	1	0.82	1	0.49
84	1	0.83	1	0.51
85	1	0.84	1	0.53
86	1	0.85	1	0.56
87	1	0.86	1	0.59
88	1	0.87	1	0.62
89	1	0.88	1	0.63
90	1	0.89	1	0.66
91	1	0.9	1	0.69
92	1	0.91	1	0.72
93	1	0.92	1	0.75
94	1	0.93	1	0.78
95	1	0.94	1	0.81
96	1	0.95	1	0.85
97	1	0.96	1	0.88
98	1	0.97	1	0.91
99	1	0.98	1	0.94
100	1	0.99	1	0.97
101	1	1	1	1

Key:

MC: Moderately Conditional

LA: Linear Agents

QK: Quasi-Kantians

HCC: Highly Conditional Cooperators